university of groningen

University Medical Center Groningen

University of Groningen

Can clade age alone explain the relationship between body size and diversity?

Etienne, Rampal S.; de Visser, Sara N.; Janzen, Thijs; Olsen, Jeanine L.; Olff, Han; Rosindell, James

Published in:
Interface Focus

DOI:
10.1098/rsfs.2011.0075

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2012

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Etienne, R. S., de Visser, S. N., Janzen, T., Olsen, J. L., Olff, H., & Rosindell, J. (2012). Can clade age alone explain the relationship between body size and diversity? *Interface Focus*, *2*(2), 170-179. https://doi.org/10.1098/rsfs.2011.0075

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 06-05-2023

Supplementary material

Protocol Data Selection

1. The kingdom of Animalia (Metazoa) is the focus of this study.
2. With the online classification of life by J. R. Anderson (http://www.gpc.edu/~janderso/historic/labman/nclasslf.htm) we compiled a list of all classified phyla, classes and orders within the Metazoa.
3. Data were collected on the following quantities:
    a. Number of extant species per family
    b. Number of extinct species per family
    c. Taxon age per family (diversification time)
    d. Body mass per extant species averaged per family

Round 1. Preliminary survey

We conducted a preliminary survey to search for taxa to include in our analysis. Taxa for which we could not find any species diversity data, or taxa that might give problems in determining individuals (*e.g.* Porifera: sponges, and Anthozoa: coral reefs) were omitted. This resulted in the phyla Bryozoa, Mollusca, Echinodermata, Chordata, and Arthropoda (encyclopedia: Walker's Mammals of the World – Nowak & Paradiso 1983; referenced sources online)

Round 2. Detailed survey

Each family resulting from our preliminary survey was examined in detail: presence of the quantities a, c and d determined whether we kept a family in the database (about one third of the families were thus removed), as these are essential to study the relationship between diversification and body size.

1. **Taxon age**: We searched by family name in the Web of Science (WoS) together with any of the terms 'speciation' – 'extinction' – 'rate*' – 'diversification' – 'taxon age' – 'fossil record' – 'phylogeny' – 'molecular clock' – 'molecular phylogenetic tree' – 'evolutionary rate' – 'phyletic evolution'. The same search was done using Scholar.Google and Google to find papers not listed in WoS.
    i. In the case of taxa for which both fossil and molecular phylogenetic data were available, the larger of the

two was used because the fossil record is likely to be an underestimate.

ii. If only a period was mentioned (*e.g.* Miocene or End-Miocene), we used the average time of this period (*e.g.*, for End-Miocene, the average of the last one third of the Miocene).

2. **Number of extant species per family**: Family diversity data were taken from encyclopaedias but adjusted with recent published material (*e.g.* African elephant), resulting from searching the literature with the family name and any of the search terms 'species' – 'diversity' – 'richness' – 'species number'.

i. Subspecies were not counted as species.

ii. The most recent estimates were chosen.

3. **Body mass**: Special effort was spent on determining the adult body masses of each species of each family (after which the average per family was taken) by contacting specialists on these taxa and several museums.

i. In the case of lesser known taxa we used encyclopaedias as well as expert communications and extrapolations (equations used from literature for conversions between length and weight).

ii. Adult body mass of extant species were averaged per family or per order. Body masses of extinct species were not taken into account, as these are only known for a limited size-biased group of vertebrate taxa.

4. **Number of extinct species per family**: The number of extinct species was found in the same source as the taxon age, or (in most cases) extracted from the online databases BioLib (http://www.biolib.cz/en/main/) and Mikko's Phylogeny Database (http://www.fmnh.helsinki.fi/users/haaramo/), encyclopaedias and taxa-specific books and the IUCN Red List for recent extinctions. This search was performed using both the family name and each of the names of all the genera within the family, because sometimes the number of extinct species in a genus rather than in a family is reported.

i. The IUCN Red List was used for the most recent extinctions.

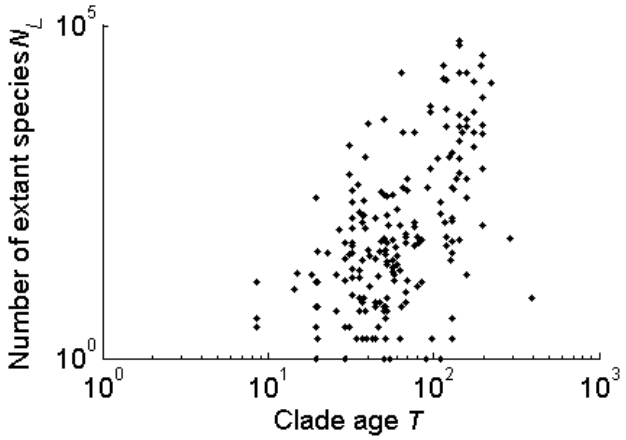
ii. Unsuccessful searches for families were taken as missing values, not as zero extinctions.

iii. Notes on newly found fossils were added to the current data.

iv. Data from insects were left out, as few fossil remains are found from these species-rich taxa and might therefore result in a clear distortion of the extinction rate.

The resulting data set is provided in the supplementary data file.

While the main text shows extant diversity versus body size and clade age versus body size, we here show extant diversity versus clade age. There is clearly a positive relationship between extant diversity and clade age. We corrected for this relationship in our analysis.



532

533 **Figure S1**. Relationship between extant diversity of a clade and the clade age.

Construction of the supertree to compute the covariance matrix

534 To calculate the covariance matrix, we created a full supertree resembling the ancestral history of all families
535 included in our dataset. The creation of the supertree was two-fold: first we determined the topology of the tree,
536 and second we determined branch lengths.

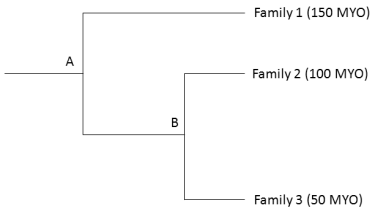
Topology

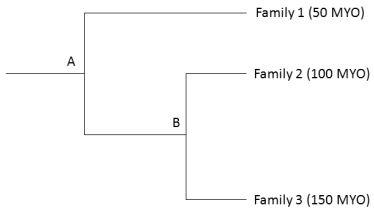
537 To create the supertree, we treated every taxonomic level as independent. Then the supertree can be split into
538 numerous sub-trees, linking all families within an order together, all orders within a class, all classes within a
539 phylum, or all phyla together. These subtrees were based on the literature (mostly molecular trees). References
540 are in the supplementary data file. We thus assume that the imposed taxonomy resembles the ancestral history of
541 the different families. We found this assumption to be robust, because if, e.g., a family would have been better
542 connected to a family outside its own order, we would have encountered this in the literature, but we did not. We
543 do note, however, that most literature used to generate the trees focuses only on a single order, class or phylum.
544 Evidently, we only need to do this for levels (e.g. order) that were non-trivially connected to the next lower level
545 (e.g. family), that is, only when more than 2 subunits occur (e.g. more than 2 families in an order).

Branch lengths

We mostly ignored the branch lengths based on molecular clocks presented in the papers used for determining the topology, because these often tend to differ substantially between papers, and we opted for a more coherent way of determining branch lengths by using the family ages estimated from the fossil record (references are in the supplementary data file). This provides a lower-limit of the age of the family, and thus the t_{ij} calculated from this supertree reflects the upper bound of shared ancestry. We illustrate our approach to determine branch lengths in Figure S1. Figure S1A shows a straight-forward topology with the fossil record data depicted on the tips. To calculate the branch lengths we have to determine the time at which the two branching events A & B happened. For simplicity we start by assigning an age to branching point B first. At branching point B two branches branch off, one leading to family 3 that first occurred in the fossil record 50 million years ago, and one to family 2 that first occurred in the fossil record 100 million years ago. We therefore assume that the younger family split off from the older family 50 million years ago, and date branching point B at 50 million years ago. In a similar fashion, branching point A gives rise to two branches, one leading to family 1 that first occurred in the fossil record 150 million years ago, and the other branch towards family 2, 100 million years old (which gives rise to family 3, including branch B, 50 million years ago). We again assume that the younger family split off from the older family, and date branching point A at 100 million years ago. We are still left with an unexplained 50 million years of the 150 million years from family 3. These are then assigned to the remaining root of the tree. Thus branch lengths in the tree are the following: Family 3 – B & Family 2 – B: 50 MY, B-A: 50 MY, Family 3 – A, 100 MY, A – root: 50 MY.

Figure S1B shows a less straight forward distribution of family ages. Similar to the previous example, we date branch B to 100 MYO. If we would proceed as before, branch A should be dated 50 MYO, because family 1 is only 50 MYA. This would however conflict with branching point B and generate an inconsistent tree. To make this tree consistent we have to assume that even though family 1 is only 50 MYO, the ancestor of this family was already present 150 million years ago. We assume here that this ancestor has either not fossilized, or his fossils have not been found yet. Thus we assign to branching point the age of 150 MY. The distance from branching point A to the root therefore remains undefined until we include more information to the tree (such as what other subtrees this subtree is connected to, or the age of the order these families belong to). Branch lengths in this subtree are thus as follows: Family 3 – B & Family 2 – B: 100 MY, B-A: 50 MY, Family 3- A: 150 MY, A-root: undefined.





573

Figure S2A Figure S2B

574 **Figure S2**. Two trees illustrating the assignment of branch lengths based on ages from the fossil record.

Calculating the covariance matrix Σ_{ij}

From the full supertree we can calculate a covariance matrix that corrects for the amount of shared ancestral time. To calculate this covariance we generate a distance matrix between all families in the tree, D. Distances are measured as the shortest distance to connect the two families, and is thus twice the distance to the closest common ancestor. Covariance between two families is than calculated as follows:

$$\Sigma_{ij} = 1 - \frac{D_{ij}}{D_{\text{max}}} \tag{16}$$

575 where D_{ij} is the distance between family i and family j, D_{max} is the maximum amount of unshared ancestral time,
576 e.g. the distance to the root of the tree. A covariance close to 1 corresponds to a high fraction of shared ancestral
577 time and recent divergence, whilst covariances close to 0 indicate that there is little shared ancestry and divergence
578 occurred long ago.

Likelihood when the allometric relationships contain noise.

579 To study the case when the allometric relationships contain noise, we took, instead of the exact relationships of
580 (10),

$$S = S_0 M^{a_S} + \varepsilon_S \tag{17a}$$

$$E = E_0 M^{a_E\text{‘}} + \varepsilon_E \tag{17b}$$

581 where ε_i ($i = S, E$) is the error. We assumed, as in all studies on allometries, that these errors are normally
582 distributed on a logarithmic scale:

$$\ln \varepsilon_i \sim N\left(0, \sigma_i^2 \Sigma\right) \tag{18}$$

where Σ is the normalized (*i.e.* scaled so that the variance components are equal to 1) variance-covariance matrix and σ_i is a scale factor. The variance-covariance matrix is determined by the phylogenetic structure of the data. We assumed Brownian motion for trait evolution (assuming speciation and extinction rates to be traits), and for Brownian motion it is well known (Martins 1995) that

$$\Sigma_{ij} \sim t_{ij} \tag{19}$$

where t_{ij} is the (relative) evolutionary time that two lineages (in this case families) have spent in common. We refer to the section Construction of the supertree to compute the covariance matrix in the Supplementary Material for more information on how we determined t_{ij}.

With the noisy allometries (17) we have to integrate over the (logarithmic) speciation and extinction rates to obtain the phylogenetically corrected loglikelihood function (where now $\Theta = \{S_0, E_0, a_S, a_E, \sigma_S, \sigma_E\}$):

$$LL = \ln \int_{\ln \overrightarrow{S}, \ln \overrightarrow{E}} \mathbb{P}\left[\overrightarrow{N}_{\mathrm{L}}(\overrightarrow{T}) | \overrightarrow{S}, \overrightarrow{E}\right] \mathbb{P}\left[\ln \overrightarrow{S}, \ln \overrightarrow{E} | \overrightarrow{M}, \Theta\right] d(\ln \overrightarrow{S})\, d(\ln \overrightarrow{E}) \tag{20}$$

where

$$\mathbb{P}\left[\overrightarrow{N}_{\mathrm{L}}(\overrightarrow{T}) | \overrightarrow{S}, \overrightarrow{E}\right] = \prod_{i=1}^{N} \mathbb{P}\left[N_{\mathrm{L},i}\left(T_i\right) | S_i, E_i; N_{\mathrm{L},i} > 0\right] \tag{21}$$

and

$$\begin{aligned} & \mathbb{P}\left[\ln \overrightarrow{S}, \ln \overrightarrow{E} | \overrightarrow{M}, \Theta\right] = \mathbb{P}\left[\ln \overrightarrow{S} | \overrightarrow{M}, \Theta\right] \mathbb{P}\left[\ln \overrightarrow{E} | \overrightarrow{M}, \Theta\right] \\ = & \frac{1}{(2\pi)^n \sqrt{|\Sigma_S|\,|\Sigma_E|}} \left(e^{-\frac{1}{2}\left(\ln \overrightarrow{S} - \overrightarrow{\mu}_S\right)^{\intercal} \Sigma_S^{-1} \left(\ln \overrightarrow{S} - \overrightarrow{\mu}_S\right)}\right) \left(e^{-\frac{1}{2}\left(\ln \overrightarrow{E} - \overrightarrow{\mu}_E\right)^{\intercal} \Sigma_E^{-1} \left(\ln \overrightarrow{E} - \overrightarrow{\mu}_E\right)}\right) \end{aligned} \tag{22}$$

with $\overrightarrow{\mu}_i$ being the mean allometries given by (10).

The integral (20) is high dimensional and therefore difficult to evaluate numerically. We used a simplex routine to find the parameter set that optimizes (20) approximated with Monte Carlo importance sampling (number of Monte Carlo samples: one million, fixed random numbers during the optimization). To check convergence we repeated this for different Monte Carlo samples (so a different set of random numbers for each optimization run), and for different starting values of the parameters.